Comparativo de Algoritmos de Machine Learning para Predição de Doença Cardíaca

Diego Juan dos Santos Silva[1], **Ericléverson Alves Ramalho de Figueirêdo**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB)

{santos.juan,ericleversoon.ramalho}@academico.ifpb.edu.br

Abstract. *The project aimed to develop a supervised predictive model to classify the presence or absence of heart disease, acting as a clinical decision support tool. Using a set of clinical and demographic data compiled from four sources (totaling 920 records), the approach was based on statistics.*

Resumo. *O projeto teve como objetivo desenvolver um modelo preditivo supervisionado para classificar a presença ou ausência de doença cardíaca, atuando como ferramenta de apoio à decisão clínica. Utilizando um conjunto de dados clínicos e demográficos compilados de quatro fontes (totalizando 920 registros), a abordagem foi fundamentada na estatística.*

1. Introdução

1.1. Contextualização do problema

O projeto se insere na intersecção entre a Tecnologia da Informação e a Cardiologia, focando no domínio das doenças cardiovasculares (DCV), que representam a principal causa de morte global. No ambiente clínico, o diagnóstico precoce é crucial para aumentar a sobrevida dos pacientes e diminuir os custos hospitalares. A triagem e o auxílio ao diagnóstico representam um desafio constante, pois os médicos lidam com uma variedade de dados heterogêneos (como glicemia, colesterol, pressão arterial e eletrocardiogramas), cuja análise manual multifatorial é complexa e suscetível à variabilidade humana.

1.2. Motivação

A principal motivação reside na necessidade de desenvolver Ferramentas de Apoio à Decisão Clínica (CDSS). O Aprendizado de Máquina é particularmente adequado para essa tarefa, pois pode identificar padrões não lineares em dados fisiológicos que são indicativos de doenças cardíacas. A relevância do problema é alta, pois visa a predição do risco cardíaco utilizando dados acessíveis e não invasivos, o que pode acelerar o diagnóstico e otimizar a alocação de recursos, evitando procedimentos onerosos em pacientes de baixo risco.

1.3. Objetivos

O problema de Machine Learning consiste em desenvolver um modelo preditivo supervisionado capaz de classificar a presença ou a ausência de doença cardíaca. O objetivo é aprender uma função de mapeamento a partir de um conjunto de dados clínicos e demográficos (como idade, colesterol e pressão arterial), que preveja uma variável alvo binária, fornecendo, assim, suporte à decisão clínica para a identificação de pacientes de alto risco. A tarefa específica de Aprendizado de Máquina é a Classificação.

1.4. Contribuições do trabalho

As contribuições deste projeto se concentram na estratégia de pré-processamento e na avaliação de modelos no contexto de saúde.

Estratégia de Tratamento de Outliers

Adotou-se uma estratégia clinicamente fundamentada, onde outliers raros (mas clinicamente possíveis, como chol =603) foram mantidos para preservar casos extremos vitais para o aprendizado do modelo. Outliers impossíveis (como chol =0 ou trestbps =0) foram tratados como dados ausentes e imputados, garantindo a integridade dos dados.

Seleção de Modelos e Otimização

A avaliação comparativa de cinco algoritmos de classificação supervisionada (Regressão Logística, Random Forest, SVM, KNN e Gradient Boosting) buscou o modelo com o melhor equilíbrio entre Precisão e Recall, sendo este último a métrica mais crítica no contexto médico (minimizar Falsos Negativos).

Interpretabilidade

O trabalho confirmou que o modelo selecionado (Gradient Boosting) capturou corretamente fatores de risco classicamente reconhecidos (dor torácica, colesterol, idade), garantindo que as predições são baseadas em sinais consistentes, o que aumenta a confiabilidade e a aplicabilidade prática do modelo.

1.5. Organização do trabalho

Preparação dos Dados

Detalha a estratégia de divisão dos dados (Hold-out 80/20 com Stratified K-Fold k=5) e o pré-processamento aplicado (tratamento de ausentes, outliers, padronização e codificação).

Modelagem e Configuração

Apresenta a seleção e a justificativa para os cinco algoritmos escolhidos, bem como o processo de otimização de hiperparâmetros (Grid Search).

Avaliação e Resultados

Compara o desempenho dos modelos usando métricas clínicas e estatísticas (Acurácia, Precisão, Recall, F1-Score e ROC), seleciona o melhor modelo (Gradient Boosting) e fornece uma análise de erros e a interpretabilidade dos resultados.

2. Trabalhos relacionados

Foram levantados dois trabalhos principais que abordam o tema de predição de doenças cardíacas utilizando Aprendizado de Máquina, servindo como base comparativa para o projeto:

2.1. Trabalhos similares

Trabalho 1: Heart disease prediction using machine learning

Problema Abordado: Foco na dificuldade do diagnóstico médico e na necessidade de sistemas automatizados, objetivando comparar diferentes algoritmos de classificação para encontrar o modelo mais preciso para a predição de doença cardíaca. Metodologia/Técnicas: O estudo comparou múltiplos algoritmos, incluindo Random Forest, SVM, K-NN, Decision Tree, ANN, Regressão Logística e Naive Bayes. O pré-processamento incluiu a remoção de outliers, tratamento de dados categóricos e escalonamento (Standard e MinMax). Principal Contribuição: Concluiu que o algoritmo Random Forest foi o de melhor desempenho, alcançando uma acurácia de 90,16%.

Trabalho 2: Heart Disease Prediction using Hybrid Machine Learning

Problema Abordado: Focado na previsão de doenças cardiovasculares, com ênfase na identificação e seleção de variáveis importantes (features) para melhorar a precisão da previsão. Metodologia/Técnicas: O estudo destacou o pré-processamento avançado, que incluiu: imputação de valores ausentes com MICE, remoção de outliers via IQR, seleção de features, escalonamento com StandardScaler e balanceamento de dados com SMOTE. O modelo proposto foi um híbrido (HRFLM) de Random Forest e Modelo Linear. Principal Contribuição: A proposta do modelo híbrido HRFLM, que alcançou 91% de acurácia no dataset Cleveland. O trabalho demonstrou que um pré-processamento avançado e a seleção de features são cruciais para a melhoria do desempenho.

2.2. Diferencial do trabalho proposto

A classificação do projeto segue a linha dos trabalhos similares, mas se diferencia pela estratégia de pré-processamento clinicamente fundamentada.

Principais Distinções

- **Tratamento de Outliers:** Enquanto outros trabalhos utilizam a remoção simples de outliers, este experimento adota uma abordagem mais robusta: mantém outliers raros (mas clinicamente possíveis, e.g., chol=603) e trata outliers impossíveis (chol=0) como dados ausentes e os imputa. Essa estratégia preserva casos extremos (pacientes graves), cruciais para o aprendizado do modelo.
- **Imputação de Ausentes:** Em vez de técnicas multivariadas complexas como MICE, o projeto utiliza a abordagem univariada mais simples e robusta, empregando a Mediana para features contínuas e a Moda para categóricas/discretas.
- **Padronização:** Foi escolhido o Z-Score, por lidar melhor com os outliers raros que foram mantidos, em contraste com o MinMax.

- **Definição do Alvo:** A variável alvo foi binarizada para 0= saudável e 1= doente (agrupando 1-4).

3. Fundamentação teórica

3.1. Conceitos básicos

O projeto utiliza o Aprendizado de Máquina Supervisionado, focando na tarefa de Classificação. O objetivo é prever uma variável alvo binária — a presença (1) ou ausência (0) de doença cardíaca — a partir de dados clínicos e demográficos.

As principais técnicas de pré-processamento empregadas foram:

- **Tratamento de Valores Ausentes:** Imputação pela Mediana (para variáveis contínuas, devido à robustez contra outliers) e pela Moda (para variáveis categóricas/discretas).
- **Tratamento de Outliers:** Manutenção de outliers clinicamente possíveis (casos extremos importantes), e tratamento de valores impossíveis (ex: chol = 0) como ausentes.
- **Codificação Categórica:** Label Encoding para features binárias e One-Hot Encoding para features nominais.
- **Padronização:** Utilização do Z-Score em features contínuas, escolhido por sua robustez aos outliers mantidos.

3.2. Algoritmos utilizados

Foram selecionados cinco algoritmos para garantir uma avaliação diversificada:

Tabela 1. Algoritmos selecionados

Algoritmo	Justificativa de Seleção
Regressão Logística	Interpretabilidade e simplicidade no domínio médico.
Random Forest	Lida bem com relações não-lineares, reduz a variância.
Support Vector Machine (SVM)	Eficaz em alta dimensão e usa kernels para problemas não-lineares.
K-Nearest Neighbors (KNN)	Captura padrões locais e serve como comparativo de similaridade.
Gradient Boosting Classifier	Constrói árvores sequencialmente, corrigindo erros para alta precisão.

3.3. Métricas de avaliação

O foco foi em métricas relevantes:

3.4. Configuração experimental

Divisão dos Dados: Estratégia Hold-out (80% Treino, 20% Teste) com Estratificação (stratify=y) para manter o equilíbrio de classes. Treino: 734 amostras. Teste: 184 amostras.

Validação: Validação Cruzada Estratificada (Stratified K-Fold) com k=5, aplicada apenas no conjunto de treinamento para otimização de hiperparâmetros e evitar overfitting.

Tabela 2. Métricas de Avaliação

Métrica	Fórmula/Definição	Relevância
Acurácia	Proporção de predições corretas.	Visão geral da eficácia (confiável devi ao balanceamento 55% vs 45%).
Recall (Sensibilidade)	VP/(VP+FN)	Métrica Crítica: Minimiza Falsos Nega vos (pacientes doentes classificados con saudáveis).
Precisão	VP/(VP+FP)	Minimiza Falsos Positivos (diagnósti desnecessário/custos).
F1-Score	Média harmônica de Precisão e Recall.	Indica o equilíbrio e robustez geral modelo.
ROC	Capacidade de distinguir entre classes.	Indica a capacidade discriminativa, ind pendente do ponto de corte.

Otimização: Grid Search foi utilizada para encontrar os melhores hiperparâmetros para cada modelo, utilizando a Acurácia como métrica de seleção primária na validação.

Recursos: Experimentos realizados em Google Colab (ambiente de nuvem).

4. Resultados e discussões

4.1. Apresentação dos resultados

Foram avaliados cinco algoritmos de classificação supervisionada (Regressão Logística, Random Forest, SVM, KNN e Gradient Boosting) no Conjunto de Teste (184 amostras). A tabela a seguir resume as métricas, sendo que o Recall (Sensibilidade) foi considerado a métrica mais crítica, pois um Falso Negativo em medicina pode ter consequências graves.

Tabela 3. Tabela de Comparação de Modelos

Modelo	Acurácia	Precisão	Recall	F1	ROC
Gradient Boosting	0.8587	0.8333	0.9314	0.8796	0.9063
Random Forest	0.8424	0.8174	0.9216	0.8664	0.9252
SVM	0.8315	0.7983	0.9314	0.8597	0.9157
Logistic Regression	0.8370	0.8214	0.9020	0.8598	0.9146
KNN	0.8098	0.7913	0.8922	0.8387	0.8890

4.2. Análise comparativa

Melhor Desempenho Geral (F1-Score): O Gradient Boosting obteve o melhor F1-Score (0.8796), indicando o melhor equilíbrio entre Precisão e Recall.

Melhor Métrica Clínica (Recall): O Gradient Boosting e o SVM empataram com o Recall máximo (0.9314).

Melhor Poder Discriminativo (AUC-ROC): O Random Forest demonstrou a maior capacidade de discriminação entre classes, com o melhor ROC (0.9252).

Modelo Selecionado: O Gradient Boosting foi escolhido como o melhor modelo por ter a acurácia mais alta (85.87%) e um Recall elevadíssimo, minimizando o risco de Falsos Negativos (classificar um paciente doente como saudável).

4.3. Discussão dos achados

1. Análise de Erros e Contexto Clínico

Todos os modelos, ao analisar a Matriz de Confusão, demonstraram uma tendência a cometer mais Falsos Positivos (classificar indivíduos saudáveis como doentes). Justificativa: Em um contexto de triagem cardíaca, este tipo de erro é o preferível. Um Falso Positivo leva a exames complementares e ansiedade, mas um Falso Negativo (dizer que um paciente doente está saudável) é mais grave, pois pode impedir o tratamento e levar a óbito. A alta taxa de Recall do Gradient Boosting reflete essa priorização, minimizando o risco de Falsos Negativos. Causas dos Erros: Os erros de classificação podem ser atribuídos à sobreposição de características fisiológicas entre pacientes (principalmente em faixas de idade, colesterol e pressão), e às limitações do dataset (ser pequeno, ruidoso e levemente desbalanceado).

2. Interpretabilidade do Melhor Modelo (Gradient Boosting)

A análise de importância de features do Gradient Boosting forneceu achados de teor exploratório para o modelo. As variáveis mais importantes foram:

Tabela 4. Feature Importance

Feature	Importância(valor)	Interpretação
cp_asymptomatic	0.2821	Variável mais importante. Está alinhada com a literatura: pacientes assintomáticos são frequentemente diagnosticados em estágios mais avançados, indicando maior probabilidade de doença significativa.
chol (Colesterol)	0.1172	O colesterol elevado é um fator de risco clássico, e o modelo identificou esta forte correlação com a doença cardíaca.
age (Idade)	0.1139	O risco de DCV aumenta com a idade, o que é clinicamente coerente.
oldpeak	0.1013	O colesterol elevado é um fator de risco clássico, e o modelo identificou esta forte correlação com a doença cardíaca.
thalch	0.0841	Frequência cardíaca máxima atingida. Frequência menor durante o esforço é comum em pacientes com limitação cardíaca.

5. Conclusão

5.1. Síntese do trabalho

O projeto desenvolveu e avaliou modelos para a classificação binária de doença cardíaca (0/1), utilizando um dataset de 918 registros clínicos. O Gradient Boosting foi selecionado como o melhor algoritmo, atingindo a Acurácia máxima (85.87%) e o melhor F1-Score

(0.8796) na comparação. Crucialmente, obteve um Recall de (0.9314), minimizando o erro clínico mais grave: classificar um paciente doente como saudável (Falso Negativo).

5.2. Principais contribuições

- **Robustez Clínica no Pré-processamento:** Ao contrário de abordagens genéricas, o pipeline manteve outliers clinicamente possíveis e corrigiu dados fisiologicamente inválidos (chol = 0) pela mediana, preservando a integridade dos casos extremos.
- **Validação Clínica da Importância:** A interpretabilidade do modelo confirmou que os preditores mais relevantes eram minimamente coerentes, como a dor torácica assintomática (cp_asymptomatic), o colesterol (chol) e a idade (age), conferindo confiabilidade ao sistema de apoio à decisão.

5.3. Limitações e Trabalhos futuros

A principal limitação foi o tamanho pequeno do dataset (menos de 1000 amostras) e a sobreposição de características que levou à ocorrência de Falsos Positivos. Como trabalhos futuros, utilizar a técnica de Data Augmentation para aumentar a quantidade e diversidade de dados disponíveis para treinar os modelos, melhorando a capacidade de generalização.

6.

Referências

Ali, S. e. a. (2024). Heart disease prediction using hybrid machine learning. *Journal of Research in Computer Science*. Disponível em: https://journal.umy.ac.id/index.php/jrc/article/view/21606. Acesso em: 11 dez. 2025.

Alotaibi, A. e. a. (2025). Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Scientific Reports*, 15. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC12006408/. Acesso em: 11 dez. 2025.

CIIS/Fiocruz (2023). Introdução à análise de dados em saúde com python. Documento técnico, CIIS/Fiocruz. Disponível em: https://docs.bvsalud.org/biblioref/2023/06/1437637/introducao-a-analise-de-dados-em-saude-com-python-ciia-saude.pdf. Acesso em: 11 dez. 2025.

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly, 3 edition. Disponível em: https://github.com/ageron/handson-ml3. Acesso em: 11 dez. 2025.

McKinney, W. (2023). *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*. O'Reilly, 3 edition. Disponível em: https://wesmckinney.com/book/. Acesso em: 11 dez. 2025.

Rahman, M. M. e. a. (2024). Machine learning approach for predicting cardiovascular diseases in low- and middle-income countries. *BMC Medical Informatics and Decision Making*, 24. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC11025260/. Acesso em: 11 dez. 2025.

Vardhan, G. H. e. a. (2022). Heart disease prediction using machine learning. *International Journal of Health Sciences*, 6(S2). Disponível em: https://sciencescholar.us/journal/index.php/ijhs/article/view/6955. Acesso em: 20 nov. 2025.

W3Schools (2025). Python machine learning.

Yaqoob, M. T. e. a. (2025). A systematic review of machine learning in heart disease prediction and diagnosis. *Frontiers in Cardiovascular Medicine*. (exemplo de revisão abrangente). Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC12614364/. Acesso em: 11 dez. 2025.